Editorial Manager(tm) for PLoS ONE
Manuscript Draft

Manuscript Number:

Title: Identification of Antifreeze Proteins and Their Functional Residues by Support Vector Machine and Genetic Algorithms based on n-Peptide Compositions

Short Title: Identify AFPs and Their Functional Residues

Article Type: Research Article

Section/Category: Other

Keywords: support vector machines; genetic algorithm; n-peptide composition; antifreeze protein; AFP

Corresponding Author: Chin Sheng Yu, Ph.D.

Corresponding Author's Institution: Feng Chia University

First Author: Chin Sheng Yu, Ph.D.

Order of Authors: Chin Sheng Yu, Ph.D.;Chih Hao Lu

Abstract: For the first time, multiple sets of global n-peptide compositions from antifreeze protein (AFP) sequences of certain cold-adapted fish and insects were analyzed using support vector machine and genetic algorithms. The identification of AFPs is difficult because they exist as evolutionarily divergent types, and because their sequences and structures are present in limited numbers in currently available databases. Our results reveal that it is feasible to identify the shared sequential features among the various structural types of AFPs. Moreover, we were able to identify residues involved in ice binding without referring to three-dimensional structures of AFPs. This approach should be useful for genomic and proteomic studies involving cold-adapted organisms.

Suggested Reviewers: Peter L. Davies
Queen's University
peter.davies@queensu.ca
expert of antifreeze protein

Brendan J J McConkey
University of Waterloo
mcconkey@uwaterloo.ca
expert of antifreeze protein recognition

Opposed Reviewers:

Dear Prof.,

Here within enclosed is our paper for consideration to be published on PloS ONE. The further information about the paper is in the following:

The Title: **Identification of antifreeze proteins and their important residues by using support vector machines based on *n*-peptide compositions**

The Authors: Chin-Sheng Yu and Chih-Hao Lu

It is first discussed that the antifreeze proteins and their functional important residues can be identified from protein sequences analysis. The common characters in antifreeze sequence still lack due to the poor homologs and radical different type in current database. Our approach not only provides excellent results for discriminating them without using the 3D structural information, but the most important, it is allowed a further investigation the rule of potential key residues in ice-binding interface.

The authors claim that none of the material in the paper has been published or is under consideration for publication elsewhere.

I am the corresponding author and my address and other information is as follows,

Address: Department of Information Engineering and Computer Science,

Feng Chia University, Taichung, 40724, Taiwan

E-mail: yucs@fcu.edu.tw

Tel: 886-4-24517250 ext. 3742

Fax: 886-4-24516101

Thank you very much for consideration!

Identification of Antifreeze Proteins and Their Functional Residues by Support Vector Machine and Genetic Algorithms based on *n*-Peptide Compositions

Chin-Sheng Yu[1,2*] and Chih-Hao Lu[3]

From the [1]Department of Information Engineering and Computer Science, [2]Master's Program in Biomedical Informatics and Biomedical Engineering, Feng Chia University, Taichung 40724, Taiwan and the [3]Graduate Institute of Molecular Systems Biomedicine, China Medical University, Taichung 40402, Taiwan

*Correspond to: Chin-Sheng Yu, Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan. FAX: +886-4-2451-6101. Phone: +886-4-2451-7250, ext. 3742. E-mail: yucs@fcu.edu.tw.

Abstract

For the first time, multiple sets of global *n*-peptide compositions from antifreeze protein (AFP) sequences of certain cold-adapted fish and insects were analyzed using support vector machine and genetic algorithms. The identification of AFPs is difficult because they exist as evolutionarily divergent types, and because their sequences and structures are present in limited numbers in currently available databases. Our results reveal that it is feasible to identify the shared sequential features among the various structural types of AFPs. Moreover, we were able to identify residues involved in ice binding without referring to three-dimensional structures of AFPs. This approach should be useful for genomic and proteomic studies involving cold-adapted organisms.

Keywords: support vector machines; genetic algorithm; *n*–peptide composition; antifreeze protein; AFP

INTRODUCTION

Antifreeze proteins (AFPs) in cold-adapted organisms prevent macroscopic ice build-up by binding to ice and thereby forestalling additional crystallization [1]. By doing so, AFPs allow organisms to survive below 0°C. It is of great interest to harness this singular property—non-antifreeze proteins cannot bind ice—for applications related to the agriculture and food industries [2,3,4,5] and to the rational design of new AFPs. However, first it is necessary to understand how AFPs and ice interact. Accurately identifying AFPs from evolutionarily divergent organisms is difficult because their sequences and structures differ radically [6,7]. To complicate matters further, for closely related species, the sequences, and consequently the structures, of their AFPs may also differ substantially if they have been geographically isolated [8]. Additionally, searching for homologous sequences within databases has not been a fruitful approach given the disparity among AFP sequences. Directly

studying AFP-ice interactions is also difficult, and a definitive picture of such interactions is not currently available [7]. Therefore, because many AFPs do not have structural or sequential features in common, it is challenging to correlate the relationships among their sequences, structures, and function.

A large number of biochemical and structural studies [9,10,11] have been performed in an attempt to understand how AFPs interact with ice on the molecular level, including site-directed mutagenesis [12,13,14] and computational experiments [15]. An ice-binding model that incorporates surface complementarity is generally accepted [16]. Recently, Doxey and colleagues [9] successfully identified AFPs, for which three-dimensional (3D) crystallographic structures were available, on the basis of their highly ordered and planar ice-binding surfaces, but their algorithm could not identify an AFP when only its NMR solution structure was available because the coordinates for the atoms at and near its surface were not well defined. [9,17]. Additionally, their algorithm requires the use of a three-dimensional (3D) structure, which is not always available for a given AFP.

It is obvious, therefore, that AFPs cannot be easily distinguished from other types of proteins. Additional information is needed to understand how AFPs and ice interact on a fundamental physicochemical level before such interactions can be applied to cold-adapted mechanisms. Although the types of amino acids present are closely coupled to the ice-binding properties of AFPs [10,13], current models usually rely on only 3D structures. To make additional use of the knowledge that has accumulated over the decades, e.g., identification of the ”hydrophobic surface” effect [7,11], the spatial regularity of an AFP solvent accessible surface, the presence of nonpolar residues, and other properties directly related to the binding properties of AFPs, an algorithm that can discern these properties is necessary. Therefore, for this report, we developed an integrated approach to rapidly identify AFPs from their amino acid sequences. Our statistically based, support vector machine (SVM) algorithm has been used to identify certain inherent protein traits e.g., protein disulfide connectivities [18], subcellular localizations [19,20], and protein folds [21], when given a query sequence, and it does not require a computational mechanical model or structure comparison. For this report, during the training and testing of this algorithm for different classifiers associated with AFPs, multiple feature schemes based on *n*-peptide compositions extracted from the sequences were used. Then, a genetic algorithm (GA) was used iteratively for key-feature selection and to improve the identification accuracy. This integrated approach enabled the recognition of AFPs on the basis of preferred short peptide sequences, rather than on structural comparisons. The identified AFP sequence features have not been reported previously, yet they correlate well with the properties of the ice-binding interfaces. This approach is suitable for the further identification of the ice-binding surfaces of AFPs.

METHODS

The Validation Dataset that Contained AFPs and non-AFPs with Known 3D Structures—

To assess our approach without bias, we tested it using a sequence validation dataset that did not

contain homologous proteins, and to examine the effects of key residues on function, we included only AFPs that had solved structures. This set contained 3762 nonredundant non-AFPs and 44 AFPs, which had been collected from the PISCES server [22] and the Protein Data Bank (PDB) [23], respectively. To include as many representative structures as possible, the non-AFPs had <25% pairwise sequence identity (SI), R-factors of 0.25 and a crystallographic resolution of at least 2 Å. The AFP sequences were separated into eight subsets on the basis of sequence identity by ClustalW2 [24]. Table 1 lists the PDB IDs of the AFPs in each subset. For a given subset, the associated AFP(s) had a sequence(s) that was not homologous to any of the AFPs in the other subsets. The non-AFPs were randomly divided among the eight subsets to cross test the performance of our approach and then were merged as a single trained model for use with other (independent) datasets (see below). Under such a critical condition, any afterward AFPs recognition so far is not simply from the self-trained sequences.

Independent Datasets—

We constructed three other datasets that did not contain the AFPs included in the aforementioned eight subsets to test our algorithm after training it with the latter. The first set included three AFP structures deposited recently in the PDB [23]; the second set contained 369 nonredundant AFP sequences deposited in the UniProKT database [25,26], which represented an evolutionarily divergent group of organisms; the third set contained two "antifreeze-like" (AFL) proteins that, while incapable of binding ice, have both a sequence and a structure that are very similar to the fish type III AFP [27]. Table 2 lists the number of AFPs derived from each type of organism included in the second dataset.

Feature schemes—

The n-peptide composition feature-based coding schemes, with n = 1 encoding the amino acid composition; n = 2, the dipeptide composition; n = 3, the tripeptide composition, etc., were used previously to predict protein properties [19,20,21,28], and we used them to characterize the important ice-binding features of AFPs. A set of symbols, A_n for the original amino acids; H_n for hydrophobicity [29]; V_n for the normalized van der Waals volume [29]; Z_n for polarizability [29]; P_n for polarity [29]; and F_n, S_n, and E_n, for groups of residues classified according to four, seven, and eight physical/chemical properties, respectively, were used to denote the feature schemes [19]. However, to characterize the key functional residues more robustly, partitioned subsequences, g-gap dipeptides, and local amino acid composition strategies were also included. [19] The partitioned amino acid composition X_k^Y is a concatenation of all amino acid sequences of composition Y and length k. The symbol D_g identifies the frequency of a sequence in the form $a(x)_g b$, where a and b denote specific amino acids and $(x)_g$ denotes the g-intervening (g-gap) residues of any type between the pair. The symbol W_l indicates the amino acid composition for peptides characterized by a set of sliding windows of length l centered on a given type of amino acid. It provides information concerning the sequential neighbors for of a given type of amino acid.

Assembly Machine-learning Algorithms—

All SVM calculations were performed using LIBSVM [30], which is a general library for support vector classification and regression, and the radial basis function kernel. In addition to the SVM algorithm [31], we implemented a GA to efficiently optimize the selection of feature attributes as detailed previously [18]. The combined use of the SVM algorithm and the GA is denoted as SVMGA. For the SVMGA, the feature attributes of each feature scheme, the penalty parameter C, the kernel parameter γ of the RBF function used for SVM identification by the GA approach were determined in advance. The GA procedure rapidly filtered out feature attributes that are not useful for SVM identification on the basis of each feature scheme.

The Voting System—

The coding scheme symbols given above denote the SVM classifiers that were derived from the various properties of the sequence descriptors. For simplicity, the participants in the SVM-identification system [19,20] were incorporated as:

$$\sum_{k=1}^{9} X_k^{A_1} + \sum_{g=0}^{6} D_g + \sum_{S} X_{k=5}^{S} + \sum_{l \in S'} W_l$$

with $S = \{H_3, V_3, Z_3, P_3, F_3, S_2, E_2\}$ and $S' = \{7, \ldots, 15\}$. The system counts the jury votes from each classifier to determine if a protein is an AFP.

Performance Assessment—

As in previous work [19,20,21], we employed the accuracy $Q_i = c_i/n_i \times 100$ to assess the performance of identification, i.e., the prediction accuracy, where c_i is the number of correctly identified AFPs in the class $i \in$ (AFP, non-AFP), and n_i is the number of sequences. The overall identification accuracy is given by

$$P = \sum_i f_i Q_i ,$$

where $f_i = n_i/N$, and N is the total number of sequences. Although Q_i provides a convenient assessment for identification performance, the Matthews Correlation Coefficient (*MCC*) [32] is a more informative measure of the performance and is given by:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} ,$$

where *TP*, *TN*, *FP*, and *FN* are the number of true positives, true negatives, false positives, and false negatives, respectively. A value for *MCC* of 1, 0, or –1 represents a perfect correlation, a random correlation, or an inverse correlation, respectively. Consideration of the *MCC*, allowed us to modify our approach to lower the number of false positives returned. To be a credible method, our approach needed to return as few false positives as possible.

AFP Sequence Homology Search—

To verify our ability to identify AFPs via their protein sequences, we tested the homology relationships among the AFP sequences. A query sequence from the second independent data set was aligned with the sequences of the 44 AFPs of the validation set. Only these 44 AFPs were used because their 3D structures have been solved, and they had been experimentally shown to bind ice. We performed an all-against-all sequence alignment using the global alignment program ALIGN [33]. Only the top-ranked sequence of the 44 AFP sequences was then used to assess the effect of homology on AFP identification, i.e., the SI value for the query sequence and the top-ranked sequence determined the usefulness of the homology search approach.

RESULTS

Identification of AFPs in a Cross-validation Dataset—

For the cross-validation test, the non-AFPs were randomly and equally divided into eight subsets, each of which contained a single representative AFP (which is identified by the first PDB ID (in bold type) in each subset list (Table 1)), and these sets formed the single representative AFP mode. Then, if the AFP representative had homologous sequences, these sequences were added into the corresponding subset. The eight subsets can be thought of as eight distant branches of an evolutionary tree. These sets formed the multiple representative AFP mode. For an experiment, the sequences of seven of the subsets were used to train the SVM algorithm with a given feature scheme, and then the output model of the trained algorithm was used to test the sequences in the subset that was not used for training. This training-and-testing cross-validation procedure was repeated eight times for a given feature scheme, each time omitting a different sequence subset during training. All results reported the performance on the total number of datasets. The SVM classifiers were optimized so that the algorithm could assign a protein sequence as either an AFP or non-AFP sequence.

Table 3 contains a summary of the identification accuracies and the MCC values for the different combinations of feature schemes used for the single representative AFP mode and the multiple representative AFP mode. Only the best result for a given feature scheme is reported. The best overall identification accuracy was 62.5% for the single representative AFP mode used by the SVM algorithm. Incorporation of the GA algorithm substantially improved the identification accuracy. Using the iterative procedures mentioned above, the GA identified the largest number of true positives and the smallest number of false positives as it discarded feature attributes that were not useful for the SVM classification. The assembled SVMGA approach correctly identified all AFPs in the single representative AFP mode. Using just the smallest possible number of selected features, the SVM classifier identified more completely structurally dissimilar AFPs than did Doxey and colleagues who used the structural characteristics of the AFPs [9]. After we decreased the number of *FP*s as much as possible (<70 FPs remained), we tested the algorithm with the multiple representative AFP mode, which was a more realistic dataset. Although the performance of the algorithm declined with the

increase in the number of divergent sequences, the identification accuracy was a respectable 54.5%.

Identification of AFPs in the Independent Datasets—

The three AFPs of the first independent dataset, which were the A chains of 2zib, 3bog, and 3boi were all accurately identified as AFPs. We observed that the sequence of 2zib is homologous to that of 2afp, which was contained in the eighth validation subset, and the sequences of 3bog and 3boi are homologous to that of 2pne, which was contained in the sixth validation subset. In addition to accurately identifying the proteins of the first independent dataset as AFPs, the algorithm also recognized that the human and bacterial AFL proteins (PDB IDs 1wvo and 1xuz, respectively) [27] were not AFPs. The human AFL and the bacterial AFL are both very similar in sequence and structure to that of the fish type III AFP (PDB code 1msi).

For the AFPs of the second independent dataset, which represent a divergent group of organisms and were collected from the UniProKT database [25,26], about 61% were correctly identified as AFPs by the SVMGA. The SI pair distribution, which characterizes the relative number of sequence pairs in the close percentage sequence identity interval, was used to examine the effect of sequence homology on AFP identification. The 369 AFP sequences were each used as a query sequence to profile the SI pair-distribution. Each query sequence was aligned with the 44 AFPs of the validation set and also with the other 368 sequences of the second independent data set. The largest SI value for each query that was aligned with the 44 AFPs was plotted along the *y* axis, and the largest SI value for corresponding sequence aligned with the other 368 sequences of the second dataset was plotted along the *x* axis (Fig. 1). The SI values associated with AFPs in the independent dataset that were incorrectly identified by the SVMGA are colored red in Figure 1, and most of these values are <20%, which below the so-called midnight-zone threshold where a structural/functional relationship can be detected [34]. Because the dataset that contained the 369 AFPs was biased as it contained AFPs from well-characterized cold-adapted organisms, many of the points were located at the far end of the *x* axis.

Coding Schemes—

For the different coding-scheme SVM classifiers used in this study, we were able to reduce the number of feature attributes required by at least 50% after implementing the GA. Consequently, each remaining classifier was well suited to identifying the corresponding type of AFP (Table 4). To understand why the features were selected as classifiers, we assigned a number (vote) when the pattern of residues in a sequence matched a GA-selected feature attribute of a coding scheme. The sequence position was marked as an SVMGA key residue if it had received a majority of the jury votes from the 14 coding schemes that we used for the multiple representative AFP mode. For instance, the dipeptide LT was selected in the D_0 scheme, and the interval dipeptide $T(X_2)T$ was selected in the D_2 scheme. Hence, for the short peptide NTALT, the L in the forth position and the first T each received one vote, and the second T received two votes (Table 5). Eight representative

AFPs are presented in Fig. 2, with their SVMGA key residues marked. Residues with >6 votes, with 4 or 5 votes, and with <3 votes are colored red, yellow, and gray, respectively. Fig. 3 illustrates the average number of SVMGA key residues in AFP sequences (black bars) and the number of in non-AFP sequences (gray bars). And it is obviously that the number of SVMGA key residues in AFP sequences is twice in non-AFPs. Approximately 70% of the SVMGA-selected key residues are solvent exposed (data not shown), which is sensible as these residues are more likely to interact with ice.

DISCUSSION

Previous studies have deduced the structural character of the interactions between ice and AFP molecules [7,14]. Knowing how ice and AFP molecules interact allows for the identification of AFPs given their structures (see the excellent results of Doxey and colleagues reported in Table 3). However, the method of Doxey and colleagues required the use of proteins with solved 3D structures, and therefore, until this report, there has not been a more general method for AFP identification.

For this report, we presented an integrated machine-learning method, SVMGA, to identify AFPs that uses multiple *n*-peptide composition features. Our results show that sequentially divergent AFPs can be identified according to their shared sequence characteristics because any test sequence or its homologs are not appearing in trained set. A set of *n*-peptide composition-based SVM predictors were combined to accurately recognize AFPs, and more importantly, to identify the key functional residues at the ice-binding surfaces. Several reports [7] have characterized defining residue repeats in AFP sequences, e.g., alanine-rich sequences in the α-helix of type I AFPs (A28–A34, Fig. 2f), and Thr-Cys-Thr (Fig. 2b) or Thr-Xaa-Thr (Fig. 2c) sequences in insect AFPs. The feature attributes, selected by our SVMGA approach, included these features. Some of the key SVMGA residues in these representative structures of AFPs, formed relatively flat planes, e.g., the red and yellow clustered regions in Fig. 2 and 4. Additionally, SVMGA approach identified some residues reside at the interface between two chains of crystallized form in PDB, e.g., T13 and T24 in chain A of 1wfa (Fig. 2f), but actually the active protein is monomer. We found others that the SVMGA key residues in red, L12, L23, A31, and T35, reside on the same side of the flat binding interface. Another similar example is the β-sheet plane of chain A in 1ezg (Fig. 2b), although the Thr-Cys-Xaa tri-peptide parallel strands [35] align perfectly in the dimer crystallized form, this flattest ice-binding surface is found in the monomer as seen by the coloration at the functional interface.

We also inspected the key residues that were identified in the eelpout type III AFP, which has been subjected to many mutagenesis studies. As mentioned in Method, this eelpout type III AFP, which PDB codes 1msi, had no homolog in any of the AFPs in trained subsets 1, 3, 4, 5, 6, 7 and 8 (Table 1.). And the key residues of 1msi were inferred from theses dissimilar trained sequences by SVMGA approach. Compared with previous studies [12,14], the SVMGA identified half of the proven ice-binding residues at the interface (Fig. 4b). For the three residues, N14, A16, and T18, which when

mutated caused the greatest decreases in AFP activity, the SVMGA method found the latter two. Although our approach failed to recognized Q9, T15, V20, and Q44, the SVMGA identified the nearby residues, L10, P12, L17, M22, V45, and V49. Residues L10 and P12 also reside at the ice-binding interface.

For the detail results obtained for the 369 AFPs in the second independent dataset (Fig. 5), for which no structural information was available, the identification accuracy diminished as the evolutionary distance of a protein sequence increased from the model fish and insect sequences. For sequences with very low SI values (15~20%), especially those from algae, bacteria, and plants, our approach was around 30% accurate. The identification of fish AFPs was around 60% accurate even when sequences with lower than 20% SI values. In fact, we believe that the features encoded in the fish and insect sequences may be used to identify AFPs from evolutionarily divergent organisms. Additionally, as more sequence data for AFPs are accumulated, they can be used to further characterize the mechanisms of cold adaptation. Finally, our approach can be used as an efficient way to obtain high throughput identification of protein function on a genome-wide scale. We have implemented iAFP web service, which is available at http://140.134.24.89/~iafp/.

ACKNOWLEDGMENTS

We thank Jenn-Kang Hwang (National Chiao Yung University) for his invaluable comments and crucial insights and Chen-Hsiung Chan (Tzu Chi University) for helpful discussions. This work was supported by grants from the National Science Council, Taiwan to CSY and the National Science Council and China Medical University, Taiwan to CHL. We are grateful for the hardware and software support by the Intelligent Digit Center at Feng Chia University and the Structural Bioinformatics Core Facility at Nation Chiao Tung University, respectively.

REFERENCES

1. Fletcher GL, Hew CL, Davies PL (2001) Antifreeze proteins of teleost fishes. Annu Rev Physiol 63: 359–390.
2. Knight CA (2000) Structural biology. Adding to the antifreeze agenda. Nature 406: 249–251.
3. Fan Y, Liu B, Wang HB, Wang SQ, Wang JF (2002) Cloning of an antifreeze protein gene in carrot and influence on freeze tolerance of transgenic tobaccos. Plant Cell Rep 21: 296–301.
4. Rubinsky B, Arav A, Devries AL (1992) The cryoprotective effect of antifreeze glycopeptides from antarctic fishes. Cryobiology 29: 69–79.

5. Griffith M, Ewart KV (1995) Antifreeze proteins and their potential use in frozen foods. Biotechnol Adv 13: 375–402.
6. Griffith M, Yaish MW (2004) Antifreeze proteins in overwintering plants: a tale of two activities. Trends Plant Sci 9: 399–405.
7. Jia Z, Davies PL (2002) Antifreeze proteins: an unusual receptor-ligand interaction. Trends Biochem Sci 27: 101–106.
8. Graham LA, Lougheed SC, Ewart KV, Davies PL (2008) Lateral transfer of a lectin-like antifreeze protein gene in fishes. PLoS ONE 3: e2616.
9. Doxey AC, Yaish MW, Griffith M, McConkey BJ (2006) Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions. Nat Biotechnol 24: 852–855.
10. Graether SP, Sykes BD (2004) Cold survival in freeze-intolerant insects: the structure and function of beta-helical antifreeze proteins. Eur J Biochem 271: 3285–3296.
11. Harding MM, Ward LG, Haymet AD (1999) Type I 'antifreeze' proteins. Structure-activity studies and mechanisms of ice growth inhibition. Eur J Biochem 264: 653–665.
12. Graether SP, DeLuca CI, Baardsnes J, Hill GA, Davies PL, et al. (1999) Quantitative and qualitative analysis of type III antifreeze protein structure and function. J Biol Chem 274: 11842–11847.
13. Graether SP, Kuiper MJ, Gagne SM, Walker VK, Jia Z, et al. (2000) Beta-helix structure and ice-binding properties of a hyperactive antifreeze protein from an insect. Nature 406: 325–328.
14. Jia Z, DeLuca CI, Chao H, Davies PL (1996) Structural basis for the binding of a globular antifreeze protein to ice. Nature 384: 285–288.
15. Nutt DR, Smith JC (2008) Dual function of the hydration layer around an antifreeze protein revealed by atomistic molecular dynamics simulations. J Am Chem Soc 130: 13066–13073.
16. Leinala EK, Davies PL, Jia Z (2002) Crystal structure of beta-helical antifreeze protein points to a general ice binding model. Structure 10: 619–627.
17. Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R (2005) Optimal docking area: a new method for predicting protein-protein interaction sites. Proteins 58: 134–143.
18. Lu CH, Chen YC, Yu CS, Hwang JK (2007) Predicting disulfide connectivity patterns. Proteins 67: 262–270.
19. Yu CS, Chen YC, Lu CH, Hwang JK (2006) Prediction of protein subcellular localization. Proteins 64: 643–651.
20. Yu CS, Lin CJ, Hwang JK (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci 13: 1402–1406.
21. Yu CS, Wang JY, Yang JM, Lyu PC, Lin CJ, et al. (2003) Fine-grained protein fold

assignment by support vector machines using generalized npeptide coding schemes and jury voting from multiple-parameter sets. Proteins 50: 531–536.

22. Wang G, Dunbrack RL, Jr. (2003) PISCES: a protein sequence culling server. Bioinformatics 19: 1589–1591.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
24. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.
25. Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28: 45–48.
26. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142–148.
27. Hamada T, Ito Y, Abe T, Hayashi F, Guntert P, et al. (2006) Solution structure of the antifreeze-like domain of human sialic acid synthase. Protein Sci 15: 1010–1016.
28. Chen YC, Hwang JK (2005) Prediction of disulfide connectivity from protein sequences. Proteins 61: 507–512.
29. Dubchak I, Muchnik I, Mayor C, Dralyuk I, Kim SH (1999) Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. Proteins 35: 401–407.
30. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. pp. Software available from http://www.csie.ntu.edu.tw/~cjlin/libsvm.
31. Vapnik V (1995) The nature of statistical learning theory. New York Springer.
32. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405: 442–451.
33. Myers EW, Miller W (1988) Optimal alignments in linear space. Comput Appl Biosci 4: 11–17.
34. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12: 85-94.
35. Liou YC, Tocilj A, Davies PL, Jia Z (2000) Mimicry of ice structure by surface hydroxyls and water of a beta-helix antifreeze protein. Nature 406: 322–324.
36. DeLano WL (2002) The PyMOL Molecular Graphics System In: Scientific. D, editor. San Carlos, CA, USA. http://www.pymol.org. .

FIGURE LEGENDS

Fig. 1. Sequence identity distribution for pairs of AFPs. The *x*-axis values are the best pairwise-matched SI values for each AFP sequence against the other 368 sequences. The *y*-axis values are the best pairwise-matched SI values for each of the 369 AFP sequences of the second independent dataset against the 44 sequences of the validation set. A black symbol indicates a correctly identified AFP in the independent data set, and a red symbol indicates an incorrectly identified AFP.

Fig. 2. Examples of key residues mapped onto the surfaces of the eight representative AFPs used in the cross-validation tests. The structures were drawn with PyMOL [36]. The residues colored in gray were not identified as key residues. The residues in red obtained more votes than did the residues in yellow. (a) 1c3y; (b) 1ezg; (c) 1eww; (d) 2pne; (e) 1c89; (f) 1wfa; (g) 2py2; (h) 2afp.

Fig. 3. Difference of the number of SVMGA key residues extracted from the 44 AFP and 3762 non-AFP sequences in cross-validation dataset, respectively. Each black bar represents the mean ± standard deviations of coverage percentage a SVMGA residue was included in a AFP sequence. Each gray bar represents for non-AFP sequence.

Fig. 4. The surface of the eelpout type III AFP (PDB ID 1msi) drawn with PyMOL [36]. (a) The key residues selected by the SVMGA are labeled in black words. Residues Q9 and N14, which were identified as key residues in a mutagenesis study but not by the SVMGA, are labeled in blue. (b) A view of the ice-binding interface, wherein all residues that are part of the interface are labeled. The residues identified by SVMGA are in red and yellow. Residues known to be important in ice binding, but not identified by the SVMGA, are in cyan. Residue I13, which was not identified by the SVMGA, is in gray. Its status as a key residue has not been determined by a mutagenesis study.

Fig. 5. The identification accuracy for the 369 AFPs from the second independent set. Each bar correlates the identification accuracy with a range of maximum SI values, which was found using the *y* axis of Figure 1 in detail ranges of SI for different species.

Table 1. The eight protein subsets used for cross-validation testing.

Subset	Type	PDB ID
1	insect AFP	**1c3y**
2	Type III fish AFP	**1c89**; 3nla; 1ucs; 1ops; 1kde; 1ame; 1msi; 1b7i; 1b7j; 1b7k; 1ekl; 1gzi; 1hg7; 1jab; 1msj; 2ame; 2jia; 2msi; 2msj; 2spg; 3ame; 3msi; 4ame; 4msi; 5msi; 6ame; 6msi; 7ame; 7msi; 8ame; 8msi; 9ame; 9msi;
3	β-helical insect AFP	**1ezg**
4	Type I fish AFP	**1wfa**; 1j5b; 1y03
5	β-helical insect AFP	**1eww**; 1l0s; 1m8n
6	insect AFP	**2pne**
7	Type II fish AFP	**2py2**
8	Type II fish AFP	**2afp**

Notes: The sequences of the PDB codes given in bold type were used for the single representative AFP mode.

Table 2. The number of antifreeze protein sequences for a given type of organism in the independent dataset that contained 369 AFPs.

Organism	Number of sequences
Algae	17
Bacteria	101
Fish	123
Insects	105
Plants	23

Table 3. The performances of SVM and SVMGA for the eight-fold cross-validation tests that used the single representative AFP mode or the multiple representative AFP mode.

		SVM		SVMGA		
Number	Subset	$C+X_3+V_3X_5$		[§]14 Feature Schemes		[§§]Doxey et al.[9]
1 (1)	1	0	(0)	1	(1)	-
1 (33)	2	0	(0)	1	(15)	(3)
1 (1)	3	1	(1)	1	(1)	(1)
1 (3)	4	0	(0)	1	(1)	(3)
1 (3)	5	1	(2)	1	(3)	(2)
1 (1)	6	1	(1)	1	(1)	-
1 (1)	7	1	(1)	1	(1)	-
1 (1)	8	1	(1)	1	(1)	(0)
AFP accuracy		62.5%	(13.6%)	100.0%	(54.5%)	(90.0%)
AFP precision		21.7%	(25.0%)	10.4%	(25.8%)	(42.9%)
Overall accuracy		99.4%	(98.5%)	98.2%	(97.7%)	(99.6%)
MCC		0.367	(0.178)	0.319	(0.365)	(0.620)
TP		5	(6)	8	(24)	(9)
TN		3744	(3744)	3693	(3693)	(3184)
FP		18	(18)	69	(69)	(12)
FN		3	(38)	0	(20)	(1)

Notes: Values given in parentheses are the number of homologous proteins accurately recognized using in the multiple representative AFP mode.

[§]14 feature schemes: $\sum_{k=1}^{3} X_k^{Al} + \sum_g D_g + \sum_S X_{k=5}^{S} + \sum_{l \in S'} W_l$ where $g = \{0,1,2,3,5\}$, $S = \{H_3, V_3, P_3, S_2,\}$, and $S' = \{9,15\}$

[§§]Doxey and colleagues used structure as the property to identify 10 AFPs in their dataset excellently. Only 2atp, for which its NMR structure was used, was not identified correctly.

Table 4. The feature schemes that enabled the recognition of the AFP in a subset when the single representative mode was used. The filled circles correlate the feature schemes with the AFPs that they identified. The AFPs are denoted according to their subsets.

	Feature Scheme									
Subset	C	W_1	D_0	D_2	D_3	S_2X_5	H_3X_5	P_3X_5	V_3X_5	Z_3X_5
1			●							
2		●	●			●				
3	●	●	●	●	●	●			●	●
4			●				●	●	●	
5	●	●		●	●	●	●	●	●	
6							●	●	●	●
7	●	●	●		●		●			
8	●	●	●			●	●	●		

Table 5. An example of votes acquired by residues in a sequence from 1msi.

Sequence		…..	Q^{9}	L^{10}	I^{11}	P^{12}	I^{13}	N^{14}	T^{15}	A^{16}	L^{17}	T^{18}	…..
Coding	C		*	*		*					*		
	X_2					*			*	*		*	
	X_3				*		*						
	D_0										*	*	
	D_1			*									
	D_2								*			*	
	D_3												
	D_5											*	
	O_3X_5									*	*	*	
	P_3X_5								*	*	*		
	V_3X_5			*	*	*				*	*	*	
	S_2X_5												
	W_9			*	*	*	*	*		*	*		
	W_{15}												
Votes		…..	1	4	3	4	2	1	3	5	6	6	…..

Figure 1

Click here to download high resolution image

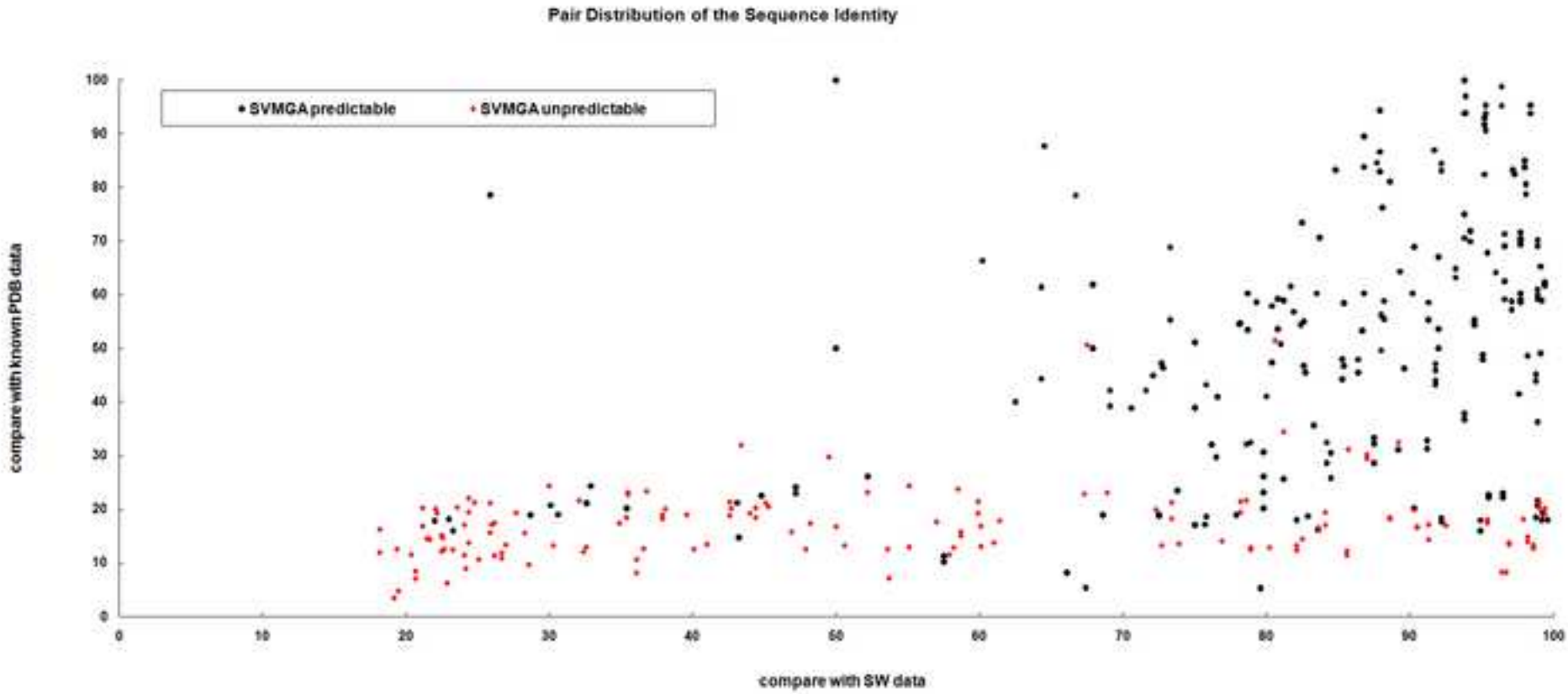


Figure 2

Click here to download high resolution image

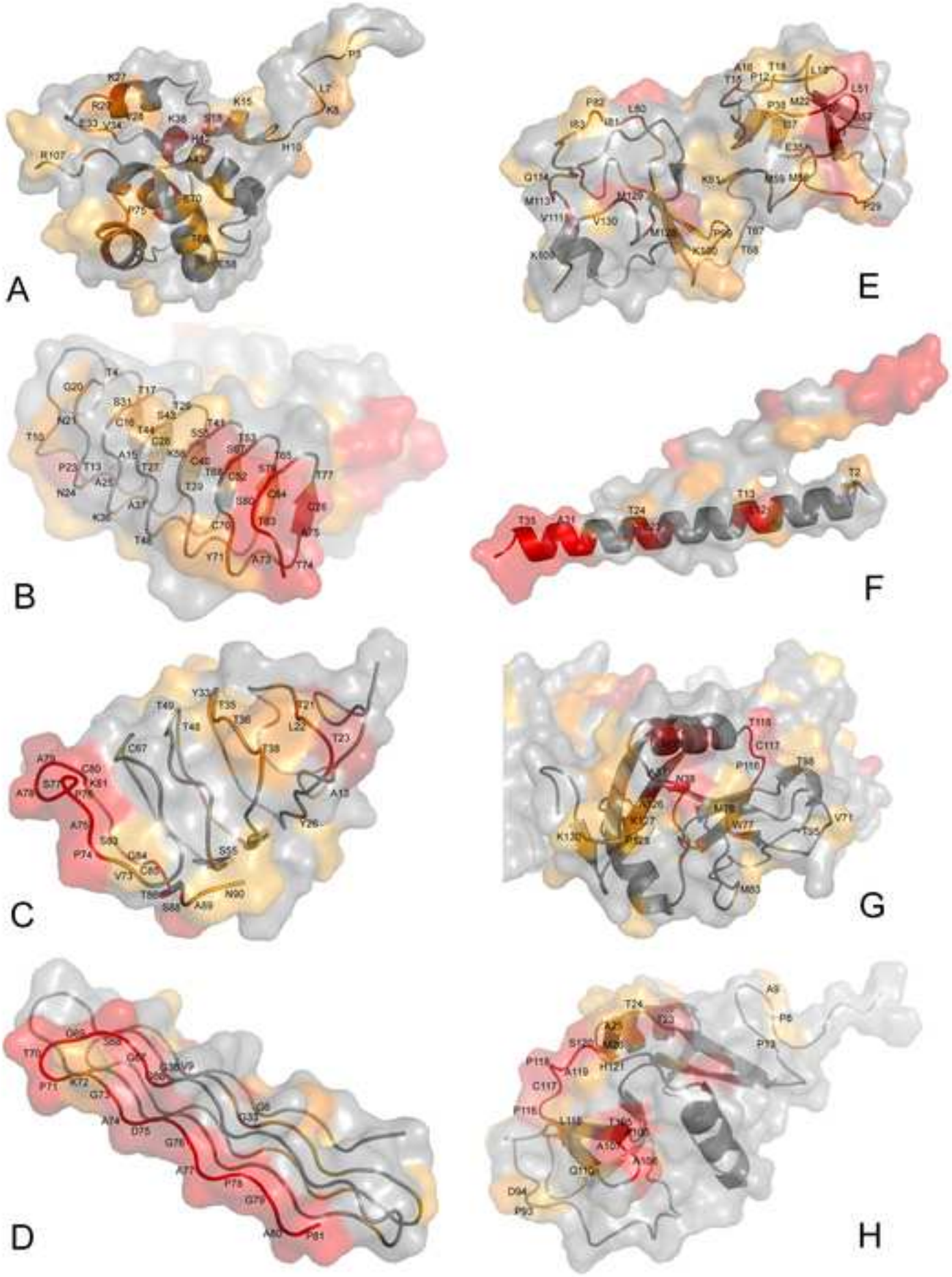


Figure 3

Click here to download high resolution image

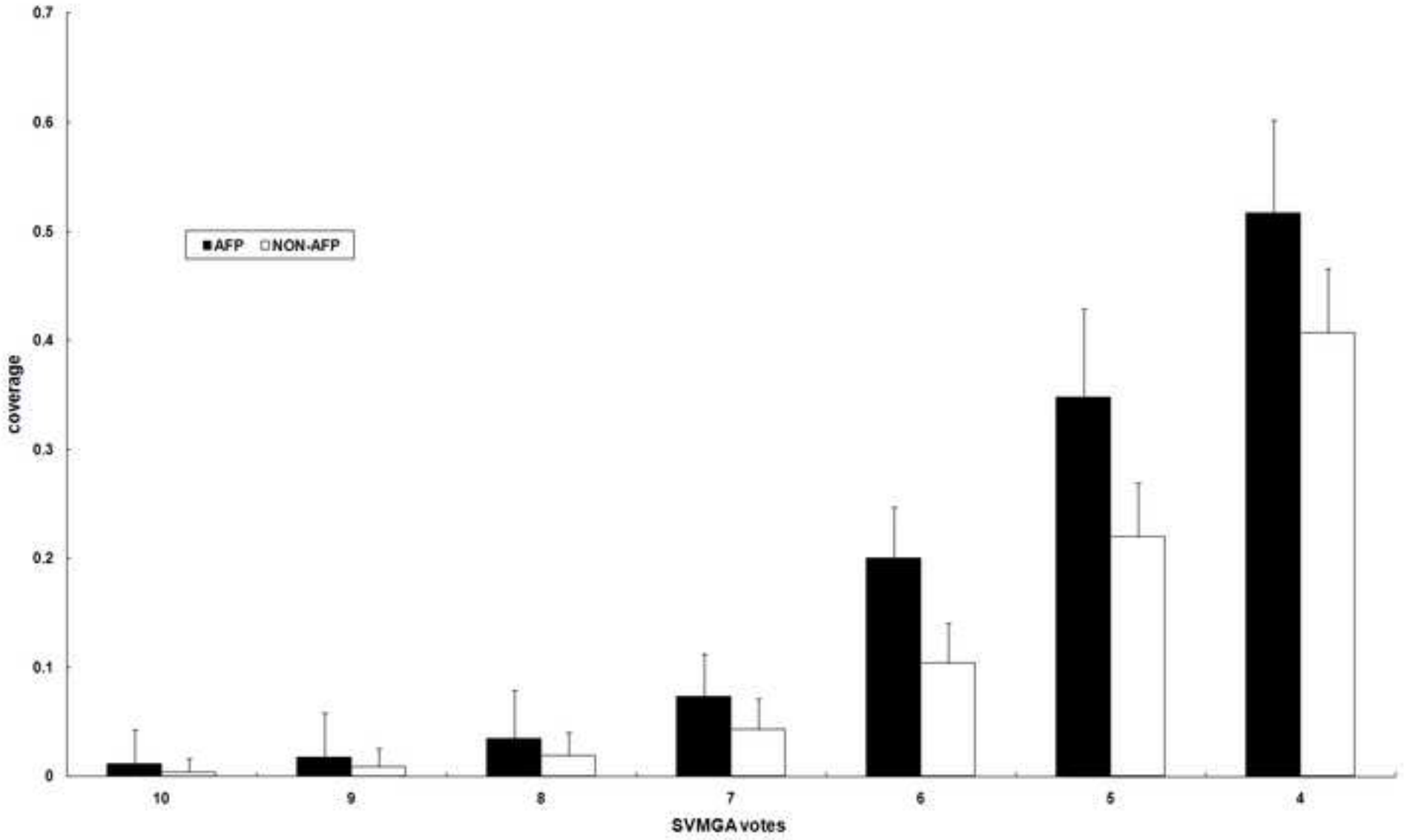


Figure 4

Click here to download high resolution image

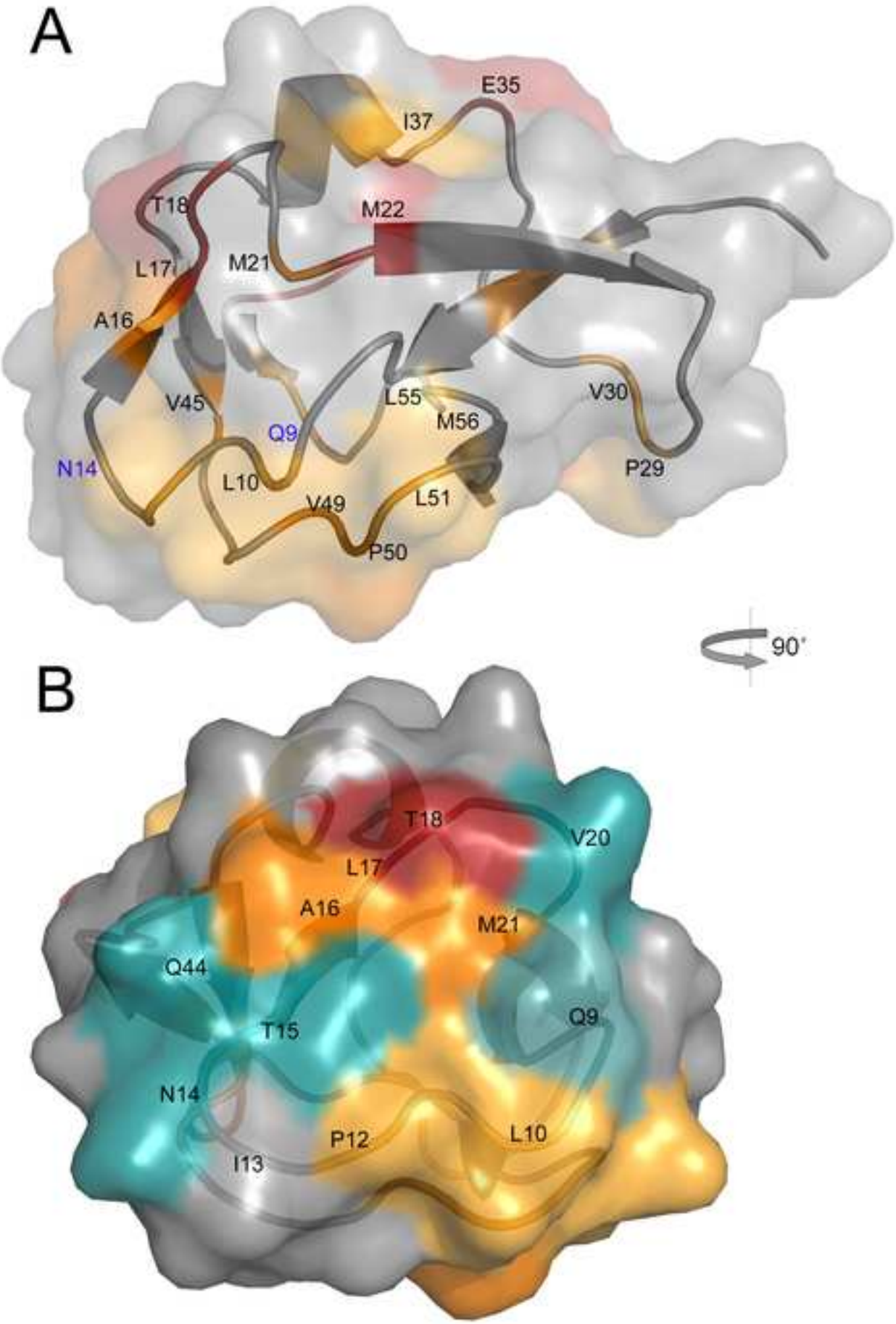
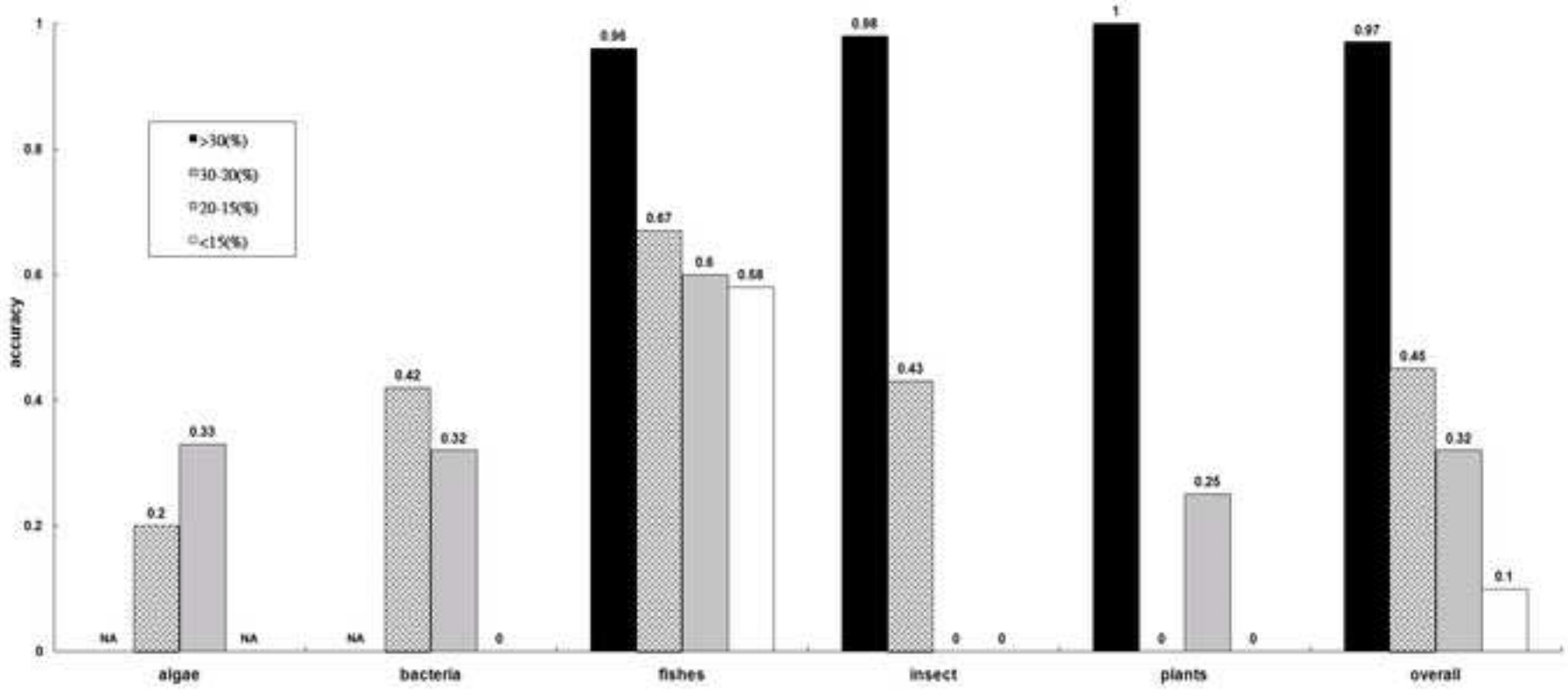


Figure 5

Click here to download high resolution image



independent dataset - 369 AFP list

Click here to download Supporting Information: AFP369SW.pdf